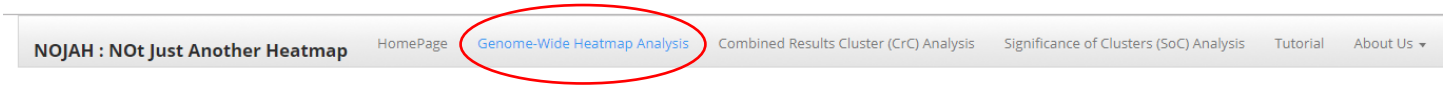
Analyzing Genomic Data with NOJAH

TAB A) GENOME WIDE ANALYSIS



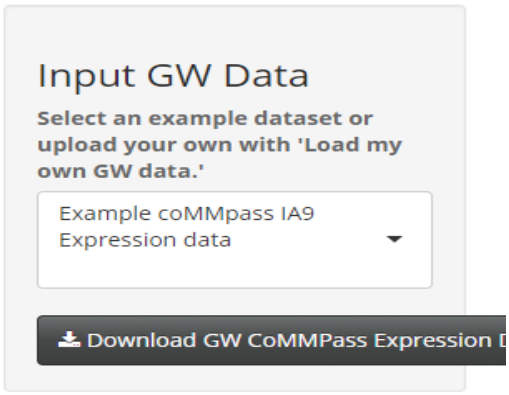
A Genome-Wide Heatmap can be very dense. Given the limitation with the computational power required to construct a genome wide heatmap, NOJAH showcases a *Genome-Wide Dendrogram.*

Genome-Wide Heatmap Analysis workflow is divided into four main subparts:

1. *Identify the Most Variable Features*
2. *Construct a HeatMap for the Most Variable Features*
3. *Identify Number of Clusters and Assess Cluster Stability*
4. *Identify Core Samples*

Heatmap is *updated* based on the Consensus Core Samples.

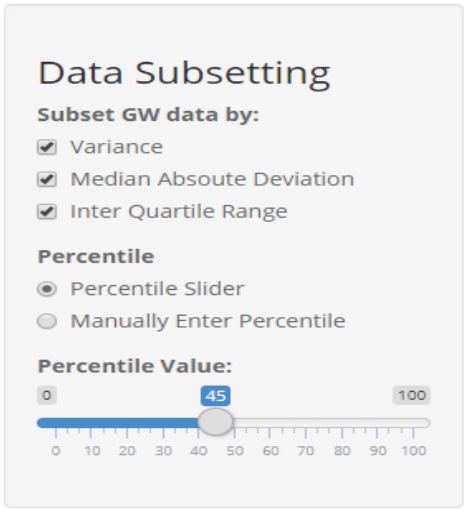
However each of these components are not dependent on each other and can be used independently.



Step 1:

Select the example dataset or upload your own. Two example datasets are available. Genome-Wide CoMMpass RNASeq Expression dataset and TCGA BRCA Expression datasets are available.

To view example data and format, use download button to view .csv file.



Step 2:

Select method of sub-setting. You can use the boxplot on the main panel to help choose the method. In the CoMMpass RNASeq example data, all three VAR, MAD and IQR show relatively larger spread.

Step 3:

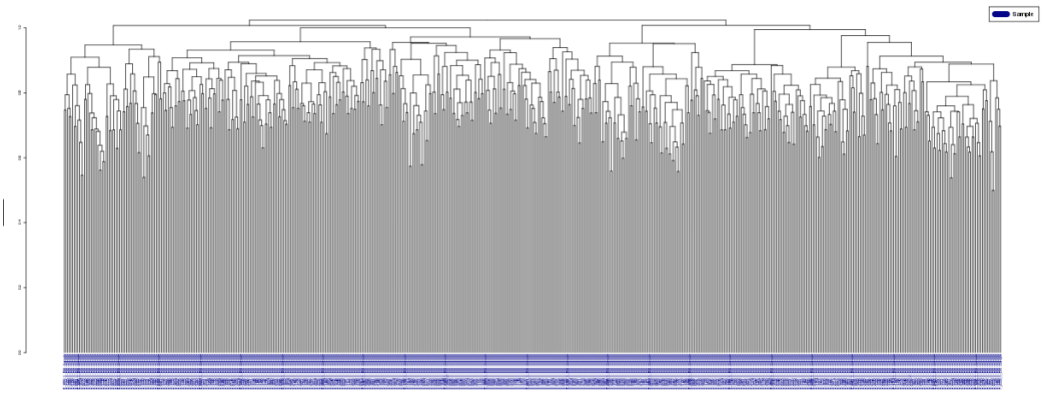
Choose a percentile cut-off to select the top most variable number of rows (genes in this case). You can also choose cut off based on the inclusion of a particular gene.

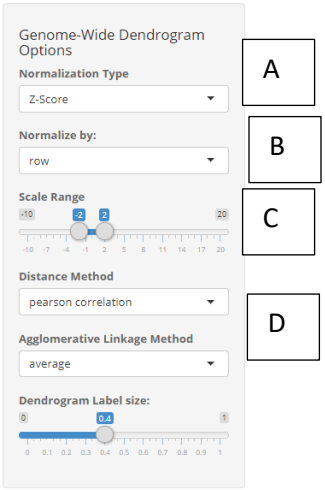
The total number of selected genes are displayed in the main panel.

Genome- Wide analysis workflow is divided into 4 subparts: (1) Identify most variable features (2) Construct a Heatmap of most variable features, (3) Identify Number of clusters and Assess Cluster Stability and (4) Identify Core Samples. Based on the core samples, heatmap is updated. User can use hyperlinks to navigate to each section. Each feature can be used individually with the own user defined data. A Genome wide dendrogram can be quite dense and sometimes not necessarily informative. NOJAH displays an Interactive Genome-Wide dendrogram instead based on different normalization and scaling methods. User can also choose between the eight different distance and seven different agglomerative linkage methods (Description on page #3).

Most-Variable Features | Heatmap | Cluster Number and Stability | Core Samples | Updated Heatmap

Genome-Wide Dendrogram





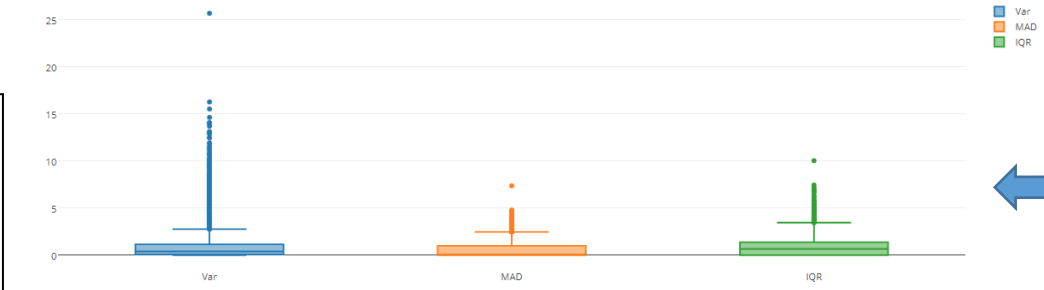
A

B

C

D

Measures of Spread

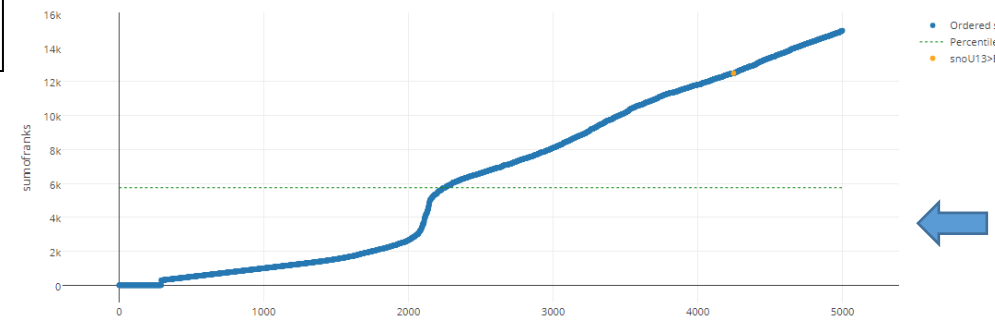


These row-wise Variance, MAD and IQR boxplots should result if you used the GW CoMMpass example data set.

Identify Most Variable Features

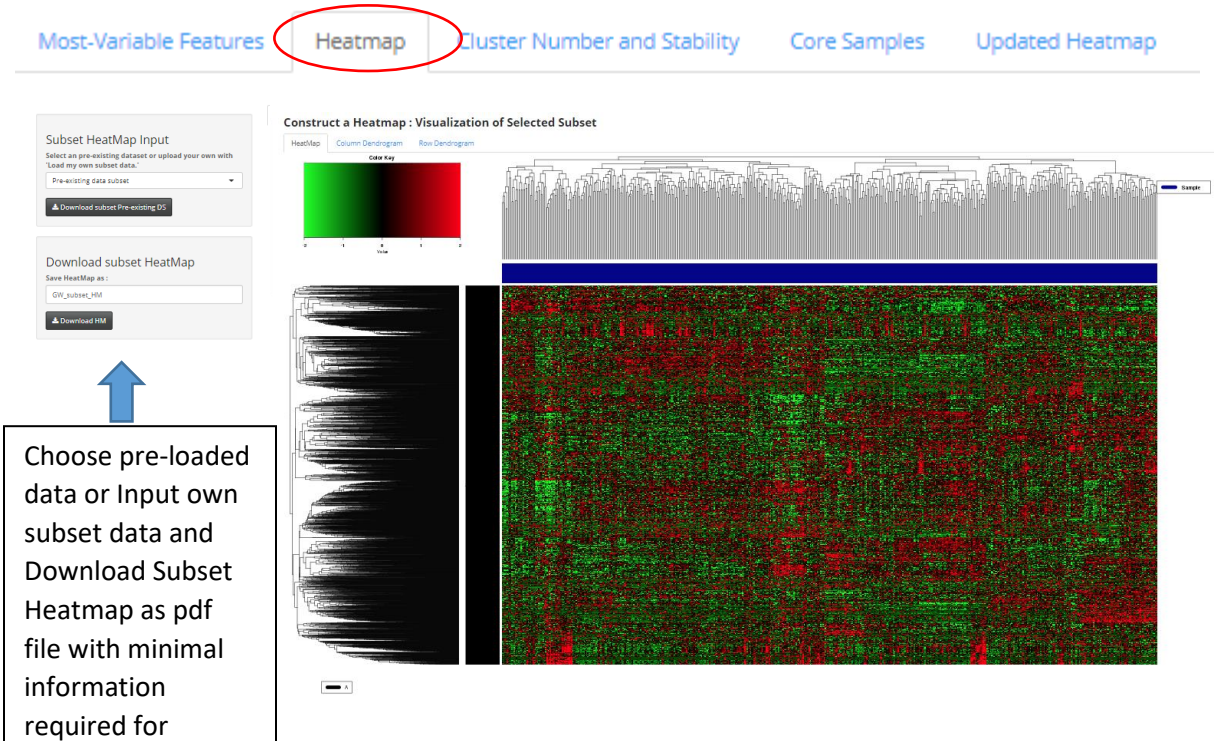
Identify Most Variable Features

To see the position of your 'gene of interest', use the 'Choose Option' drop down to the right



The number of genes selected : **2750**

Most variable genes selected based on a percentile or row inclusion criteria. Specific gene can be chosen from drop-down and highlighted in orange on plot. The number of genes included are displayed.



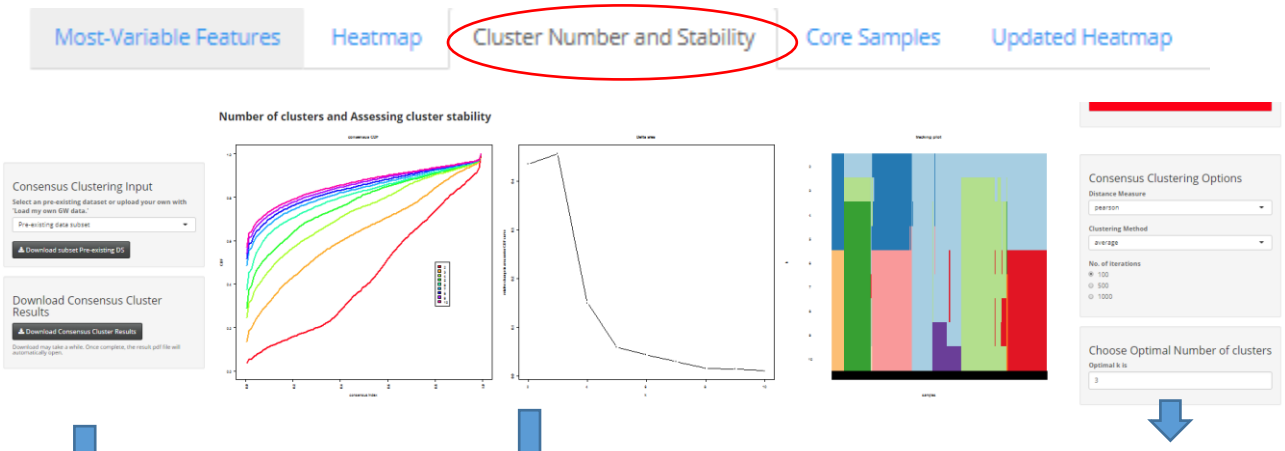
Construct a heatmap with most variable features

Choose pre-loaded data or Input own subset data and Download Subset Heatmap as pdf file with minimal information required for reproducibility. User can choose file name.

Interactive HeatMap of the top most variable genes selected using the above criteria. Separate tabs display column and row dendograms. (See tutorial for part C for detailed run-through of heatmap options, Page #9)

A. Data is z-scored before input into the heatmap.2 function.
B. Data can be normalized by row, column or both
C. Scale is set from -2 to 2 but can be changed by the user
D. Choose clustering and distance measure
E. Supervised row-wise or column wise clustering can be selected using FALSE option. Row and column labels can be displayed using TRUE option
F. Change color of HeatMap using the high, mid and low colors

Identify number of clusters and Access cluster stability



Choose pre-loaded data or Input own subset data. Data format should be same as that in the preloaded example data (available for download) and Download Consensus cluster output for consensus heatmap as available from 'ConsensusCluster' Plus Bioconductor package.

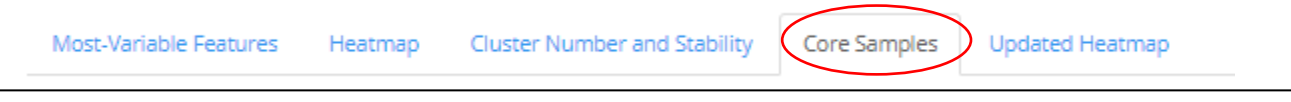
Consensus CDF, Delta area plot are from the output results of the 'ConsensusClusterPlus' package. Along with the consensus matrix heatmap (available for download), they will help the user determine the optimal number of clusters in the data.

For this CoMMPass Expression data, three optimal clusters are predicted. The user can change the optimal clusters using the right panel.

Choose distance and clustering measures to perform **consensus clustering** using the 'ConsensusClusterPlus' package, to predict optimal number Sample clusters for the data.

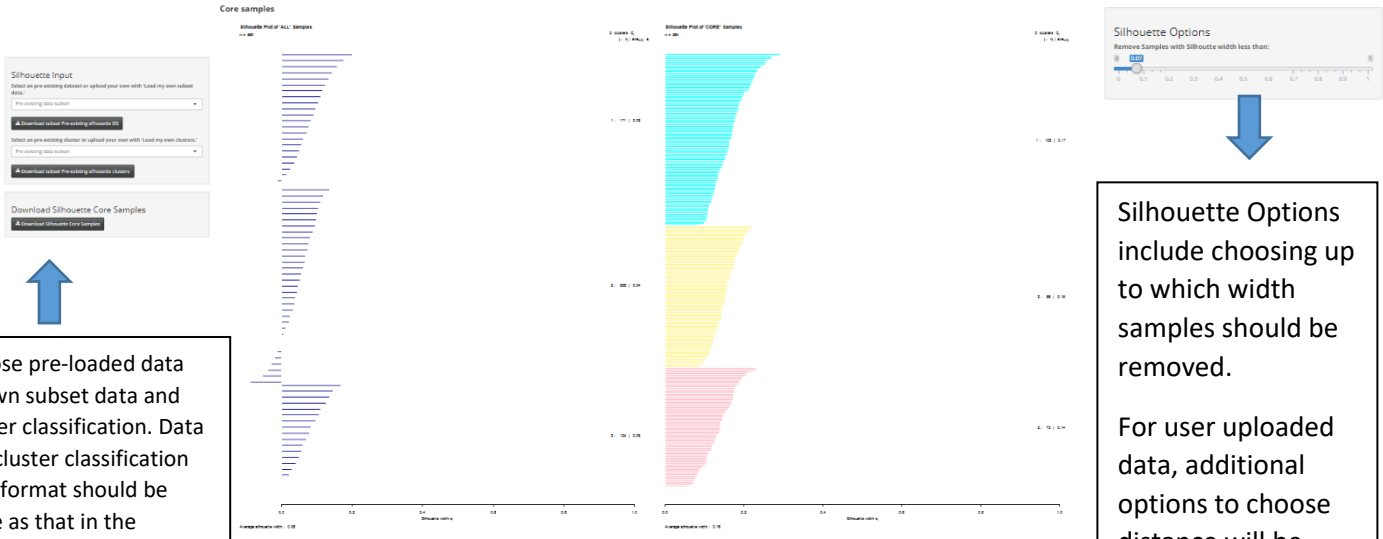
<u>Static parameter settings used:</u>

- item resampling = 80%
- gene resampling = 80 %
- maximum evaluated k = 10
- clustering algorithm = Agglomerative Hierarchical clustering algorithm "hc"
- Same clustering method is applied to both inner Linkage and final Linkage parameters.



The larger the silhouette width, the better the stability. Samples with negative silhouette width signify poor cluster stability and are removed to create the core sample set.

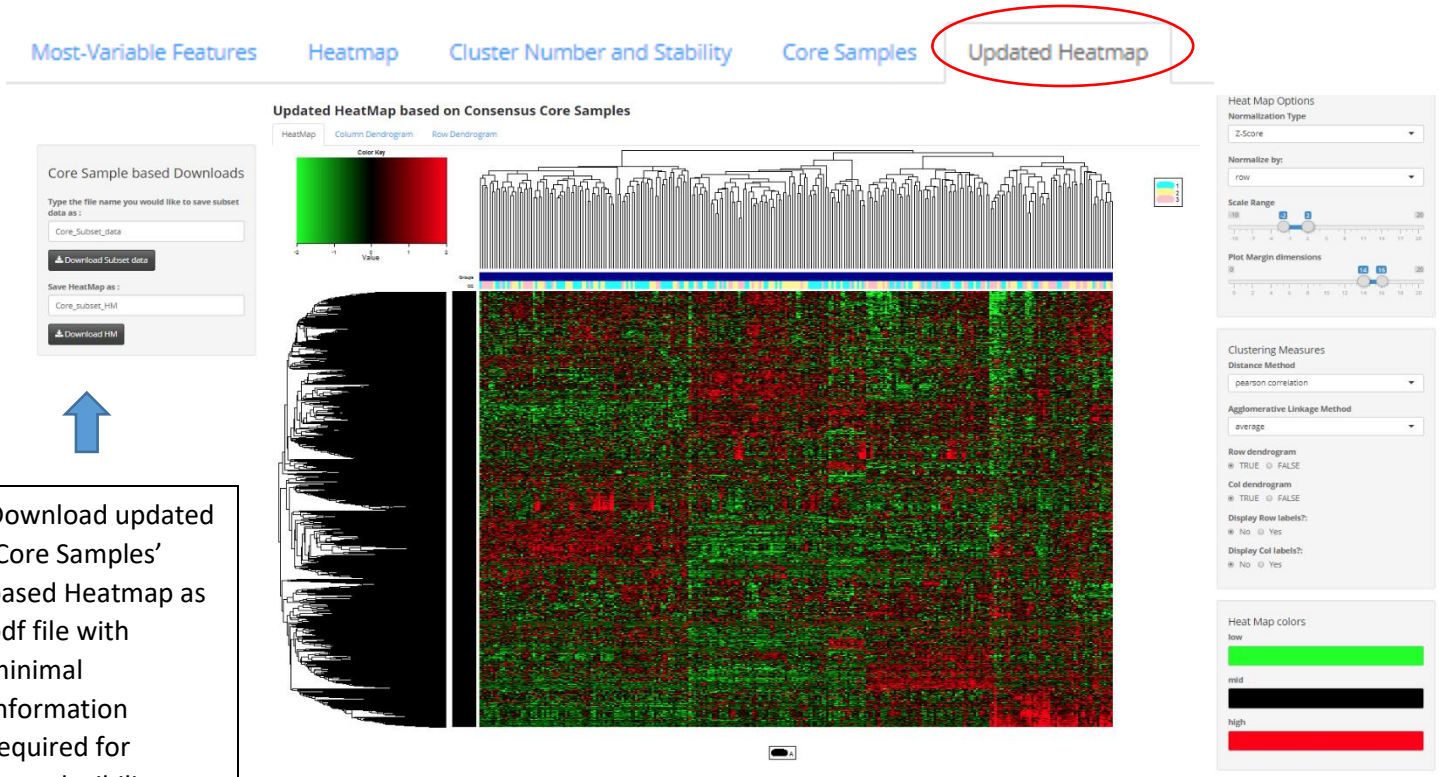
Identify Core Samples



Choose pre-loaded data or own subset data and cluster classification. Data and cluster classification data format should be same as that in the preloaded example data (available for download).

Silhouette Options include choosing up to which width samples should be removed.

For user uploaded data, additional options to choose distance will be made available.



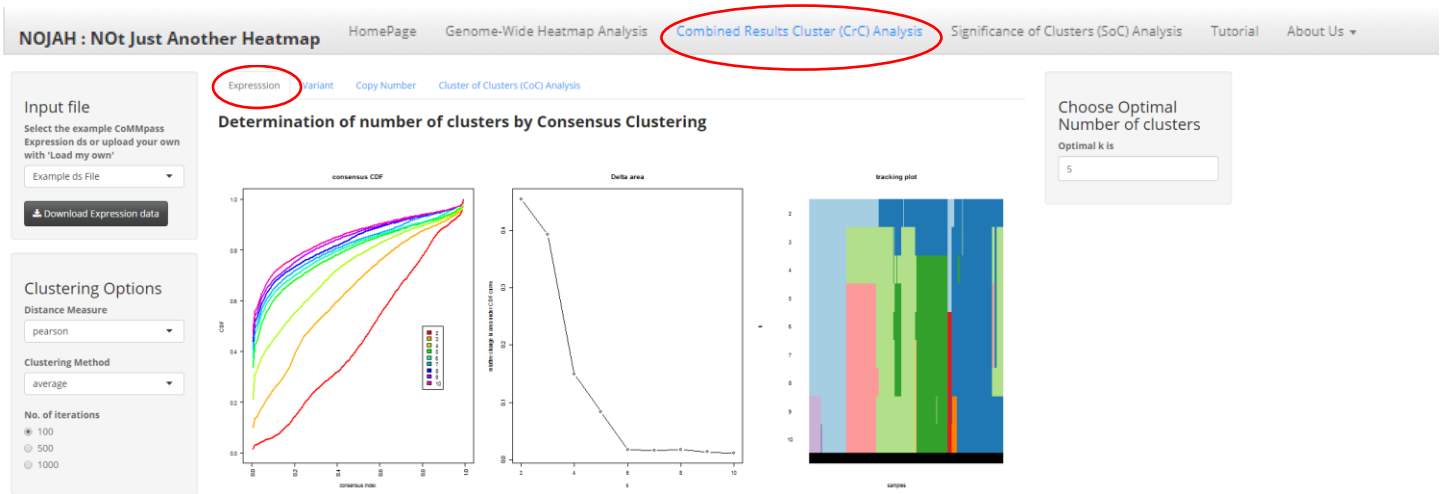
Download updated 'Core Samples' based Heatmap as pdf file with minimal information required for reproducibility. User can choose file name.

Interactive Heatmap of the top most variable genes for the Core Samples. Heatmap clustering is based on the consensus clusters (CC). The original sample clustering is available as the column color bar over the CC. Separate tabs display column and row dendrograms. (See tutorial for part C for detailed run-through of heatmap options, Page #9).

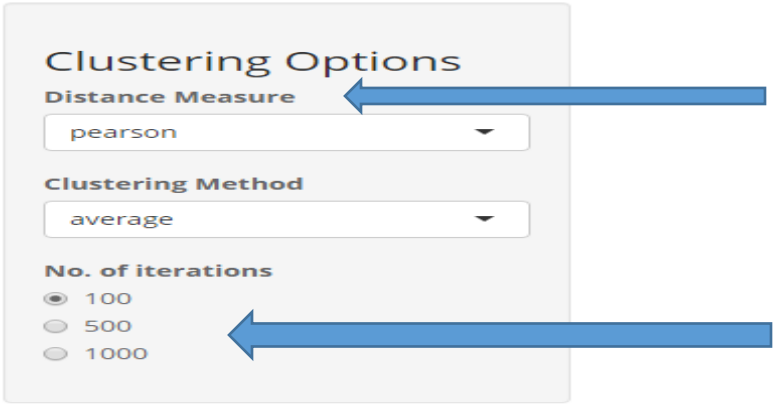
Options for heatmap are similar as shown on Page #3.

TAB B) CoMMpass DATA ANALYSIS

**Note: The same steps apply to Expression, Variant and Copy Number tabs.*



Step 1:
Select a pre-filtered RNASeq Expression CoMMpass IA9 data file or input your own coMMpass expression file.

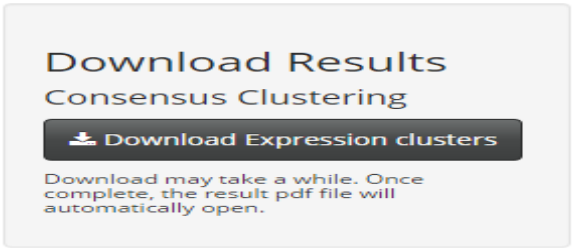


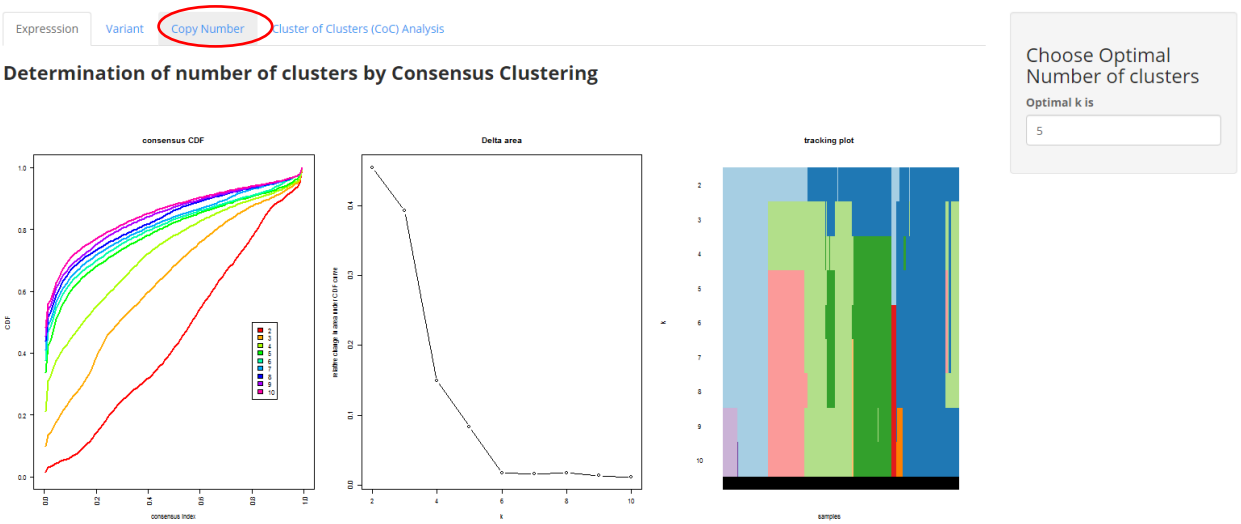
Step 2:
Choose distance and clustering measures to perform **consensus clustering** using the 'ConsensusClusterPlus' Bioconductor package, to predict optimal number of Sample clusters for the data.

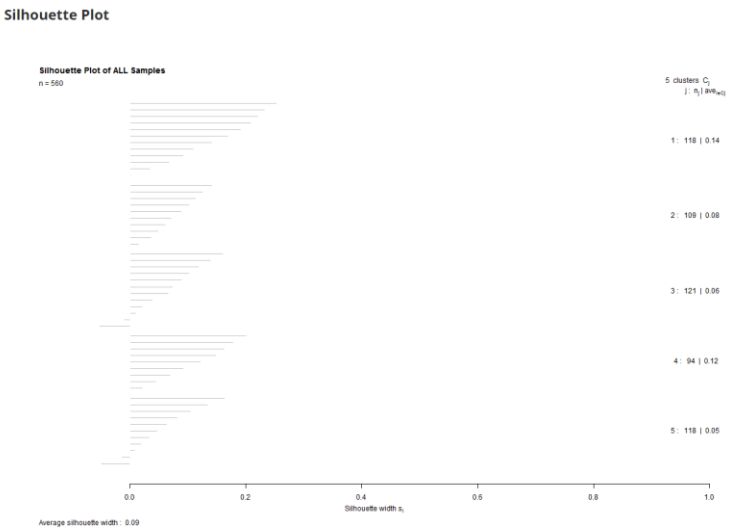
Step 3:
Choose the no. of iterations. Default is set to 100 for faster computing. In practice, set this to 1000 iterations.

<u>Static parameter settings used:</u>

- item resampling = 80%
- gene resampling = 80 %
- maximum evaluated k = 10
- clustering algorithm = Agglomerative Hierarchical clustering algorithm "hc"
- Same clustering method is applied to both inner Linkage and final Linkage parameters.







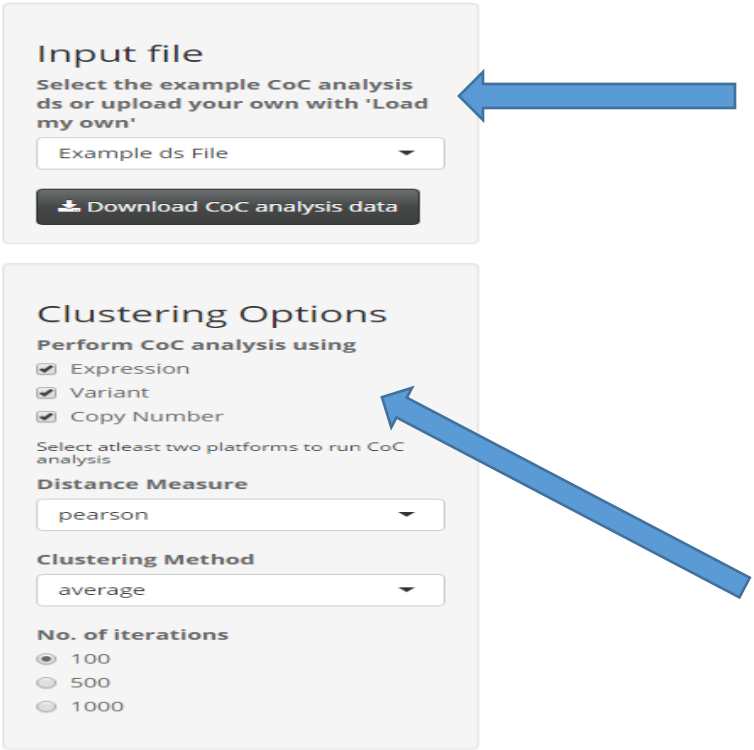
These plots will be displayed using the parameter setting above.

Consensus CDF, Delta area plot are from the output results of the 'ConsensusClusterPlus' package. Along with the consensus matrix heatmap (available for download), they will help the user determine the optimal number of clusters in the data.

For this Expression data, five optimal clusters are predicted. The user can change the optimal clusters using the right panel.

Silhouette Plot can further help confirm the identification of the number of clusters visually. The larger the average silhouette width, the more reliable the cluster structures are.

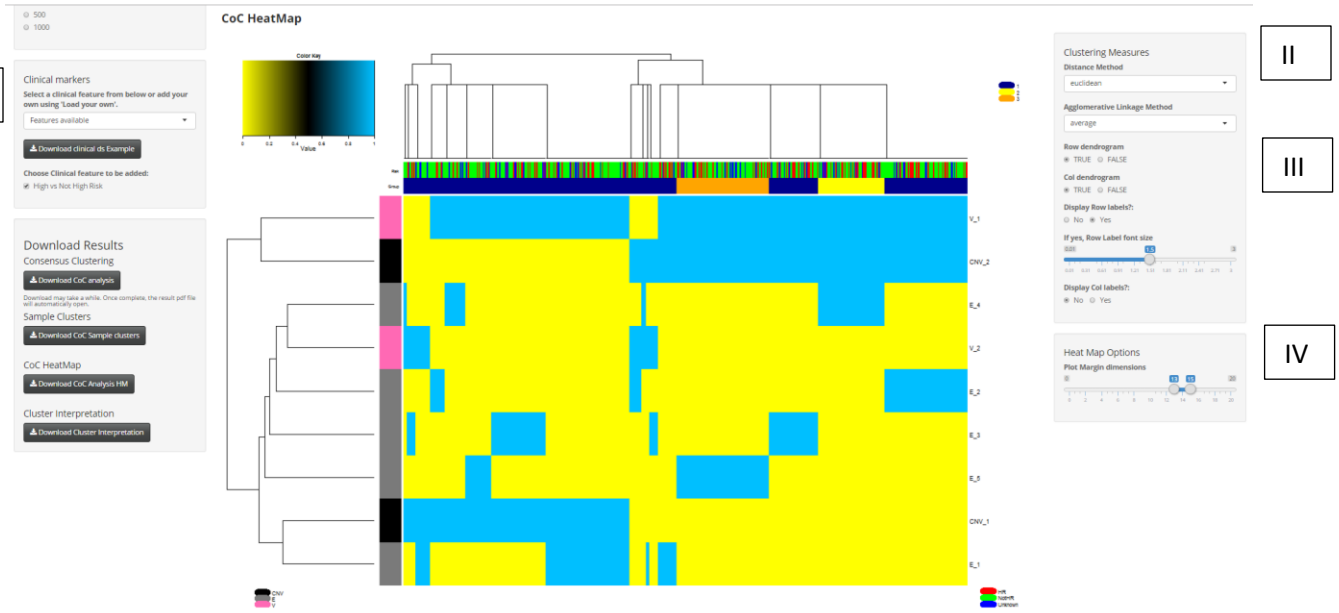
CLUSTER OF CLUSTER ANALYSIS



Step 1:
Select the computed number of optimal RNASeq Expression, Variant and CNV data from the previous 3 tabs or upload the your own clusters using the load my own option. The user-uploaded cluster data should be in the same format as the example data. Example data is available for download. In addition, the same patients should be used for each platform in order to be used for CoCA.

Step 2:
Select the platforms to base CoC analysis. User should select at least two platforms.

Select other parameter setting in the similar fashion to the previous tabs to run 'ConsensusClusterPlus' package.



Interactive HeatMap for Cluster of Cluster Analysis.

1-0 Transformed matrix data based on the individual platform clusters is used as input into the modified heatmap.2 function.

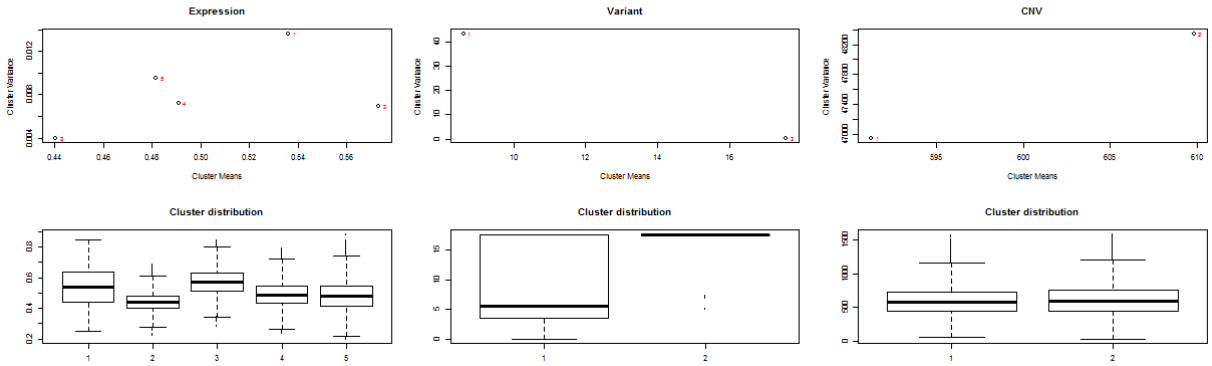
I. Add Single or multiple clinical feature(s) as bars just below the dendrogram. As an example, sample risk status is displayed above the predicted consensus cluster bar and can be downloaded using the download button.

II. Choose clustering and distance measure

III. Supervised row-wise or column wise clustering can be selected using FALSE option. Display Row and column labels using TRUE option. Adjust size of the labels using the slider.

IV. Adjust Plot margins using the slider.

Cluster Interpretation

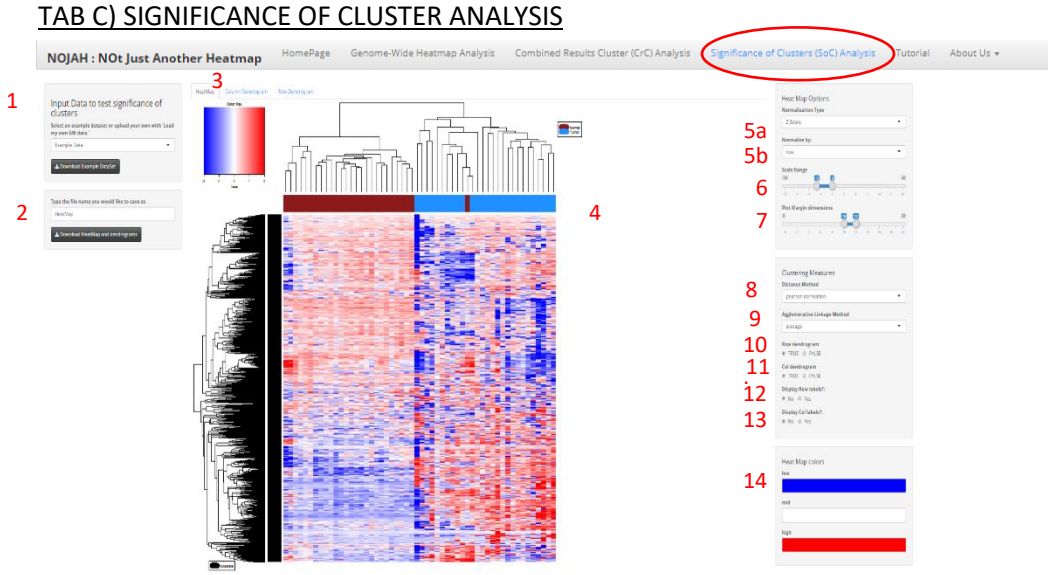


Interpretation of CoC Analysis Cluster HM based on the individual platform clusters.

Variance vs the mean plot of the lower triangular distance matrix serves as a relative measure of each cluster relative of the others within the same platform.

Boxplot of the individual clusters also helps determine which cluster has a relatively higher or lower median Expression (or median proportion or median CNV segment mean). The spread among the clusters is also informative.

TAB C) SIGNIFICANCE OF CLUSTER ANALYSIS



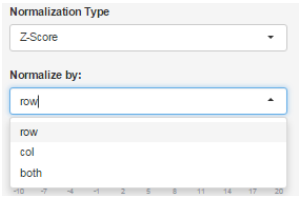
1: Select dataset of interest. Using the dropdown, you can choose the example or upload your own. If uploading your own, format data in same format as in the example file.

2: Download example data using download button to view contents/formatting of example file.

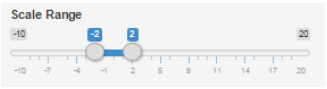
3: If example file is chosen, Heatmap automatically displayed in the HeatMap tab.

4: HeatMap created using Z-score 'row' normalization, 'Pearson correlation' distance and 'average' agglomerative linkage method (i.e. default settings). Depending on dataset may take several minutes to load.

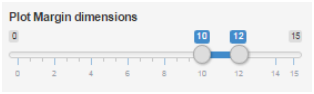
5a, b: Select a different normalization method you'd like for the data using drop down options. Each time a different type is chosen, the heatmap will be updated.



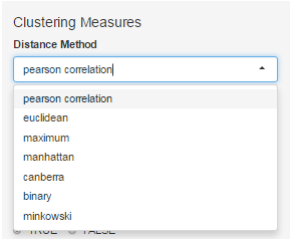
6 (optional): Drag slider to change scale range for the colors. Heatmap will be updated on movement.

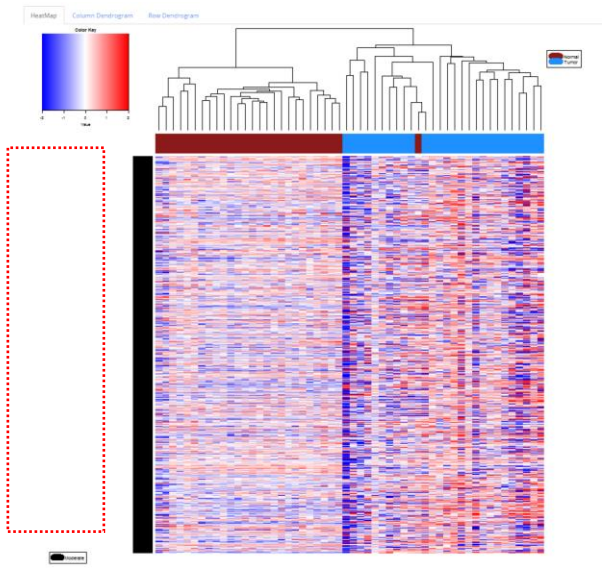


7: Select the Plot margins. If column dendrogram overlaps the legend, increase both margin points and vice – versa until desired.

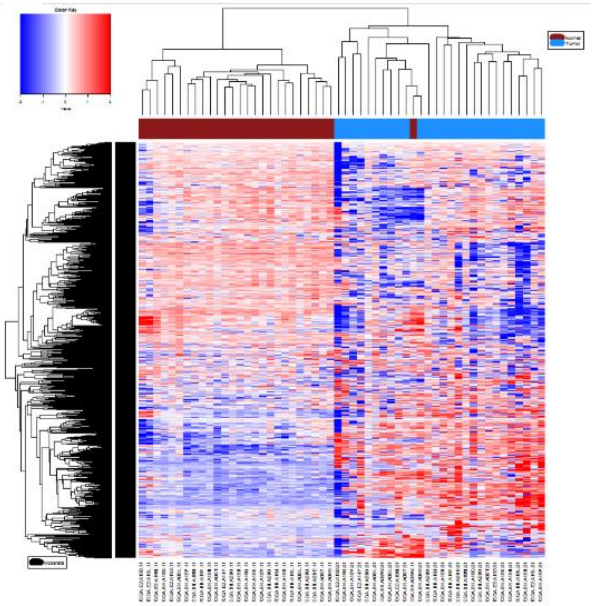


8, 9: Select Distance method and linkage method of choice using the drop down options. Each selection will display modified heatmap.

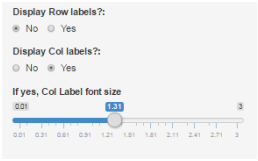


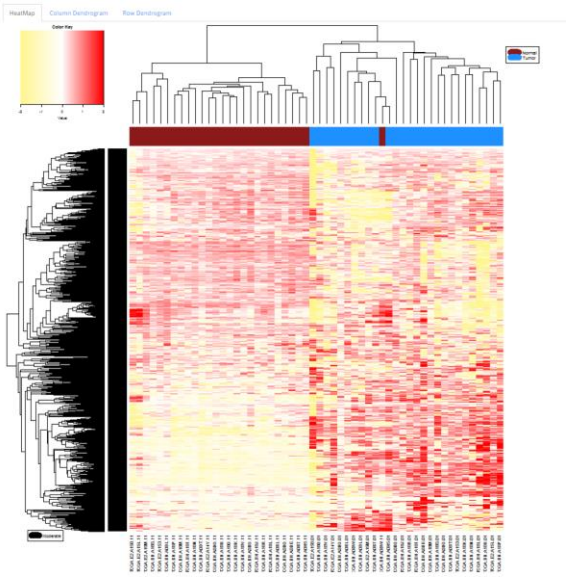


10, 11: Select either to display Row dendrogram or not. If FALSE is chosen, row dendrogram will disappear and data will not be ordered based on means. Same applies to Column dendrogram.



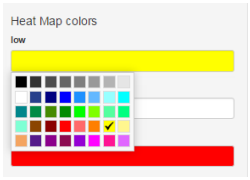
12, 13: Select Display Row labels = 'Yes' to see the corresponding CpG sites. Additional slider appears to select, font size. Same applies to Sample labels.

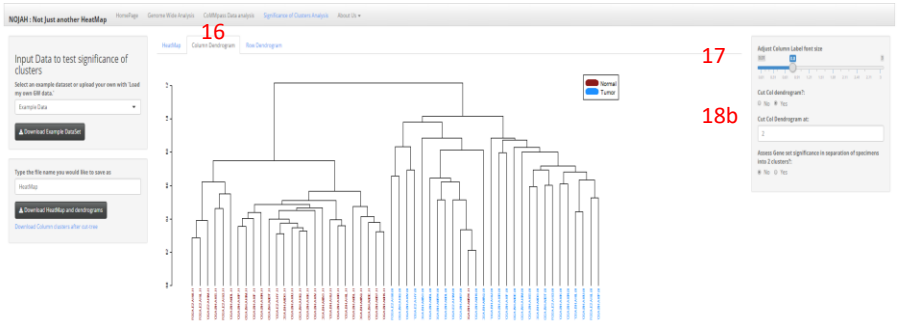




14: Select color scheme. Red-Black-Green is typically used for Expression data and Blue-White-Red is used to represent methylation data. Heatmap will update as soon as color is chosen. After choosing desired color(s), click anywhere on screen to come out of color selection panel.

15: Input file name and click on Download button to save heatmap and the corresponding row and column dendrograms in pdf format as shown below using Chrome browser.





16

17

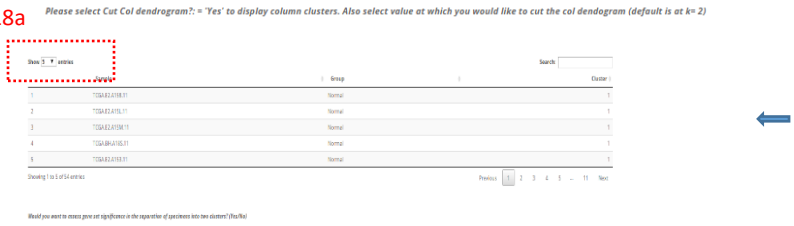
18b

20

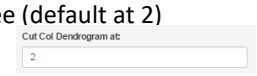
16: View in column dendrogram tab

17: Slider to adjust font size of the column dendrogram labels

18a



18 a, b: a. Option to cut the tree. b. If yes is chosen, user is asked at which position they want to cut the tree (default at 2)

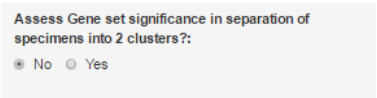


When selected, a table will appear that classifies Samples, their Groups, and their corresponding clusters.

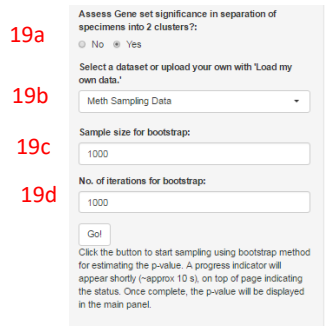
Use the drop down on upper left , to display 5/10/All rows of the table.

19 *Would you want to assess gene set significance in the separation of specimens into two clusters? (Yes/No)*

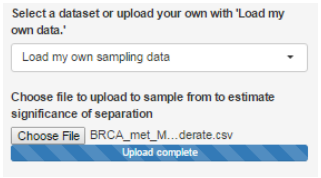
19: Option to assess gene set significance in separation of the two clusters (Tumor vs Normal). Applicable only when >=2 clusters are available for analysis.



When 'Yes' is selected, parameters for Monte Carlo p-value estimation will be made available.



19a: Select Sampling dataset for bootstrap. An example Methylation Sampling data is available or user can input their own (up to 75 MB is allowed). Large .csv and .txt files can be converted to .RDS file contain file size within 75 MB limit.

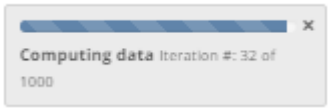


19b: Choose Sample size of the data for bootstrap. Use a size that does not exceed the original sampling data itself.

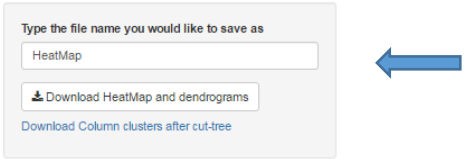
19c: Select number of iterations you wish to perform. A good practice is to perform at least 1000 iterations for accuracy of analysis.

19d: Once all options are selected, press 'Go' button to start analysis.

After approximately 10 seconds, a progress indicator will appear to track the time remaining for the analysis to be completed.



p-value results from the boot strap approach for calculation significance of clusters using Fisher's exact test will be displayed under the table along with the interpretation.



20: To download the p-value results as well, input the file names and click on Download button. The heatmap and the corresponding row and column dendrograms followed by the p-value results will be downloaded in pdf format.

To download the table for the classification of samples by clusters, click on link and the table will be saved as a .csv file.

Similar analysis can be performed on Row Dendrogram, provided you have at-least two row groups.